TITLE:

Note on robust model selection by density power divergence in a contaminated regression model (Statistical Information in Inference and Its Related Topics)

AUTHOR(S):

矢田, 和善; 青嶋, 誠; 小林, 裕子

CITATION:

矢田, 和善 ...[et al]. Note on robust model selection by density power divergence in a contaminated regression model (Statistical Information in Inference and Its Related Topics). 数理解析研究所講究録 2011, 1758: 150-159

ISSUE DATE:

2011-08

URL:

http://hdl.handle.net/2433/171314

RIGHT:

京都大学学術情報リポジトリ
KURENAI 紅
Kyoto University Research Information Repository

数理解析研究所講究録
第1758巻 2011年 150-159

Note on robust model selection by density power divergence in a contaminated regression model

筑波大学・数学系　矢田　和善 (Kazuyoshi Yata)
Institute of Mathematics
University of Tsukuba

筑波大学・数学系　青嶋　誠 (Makoto Aoshima)
Institute of Mathematics
University of Tsukuba

筑波大学大学院・数理物質科学研究科　小林　裕子 (Yuko Kobayashi)
Graduate School of Pure and Applied Sciences
University of Tsukuba

1. はじめに

パラメトリックなモデルの推定において，ダイバージェンス最小化による推定法は長い歴史をもつ．最もよく知られるのは，Kullback-Leibler情報量（K-L情報量）最小化に基づく最尤推定であろう．標本数が十分ではないとき，最尤推定量(MLE)は異常値の影響を受けやすいという欠点をもつ．一方，Hjort (1994), Scott (2001)は，L_2距離最小化に基づくL_2推定量を提案した．L_2推定量は異常値に対して頑健ではあるものの，漸近有効ではないという欠点をもつ．これらに対して，Basu et al. (1998)はdensity power divergence (DPD)を提案した．DPDは，チューニングパラメータ$\alpha(\geq 0)$の値を変えることによって頑健性と漸近有効性のトレードオフを調節した推定を考えることができ，$\alpha = 0$のときにK-L情報量と一致し，$\alpha = 1$のときにL_2距離と一致する．Jones et al. (2001), Basu et al. (2006)は，DPD最小化によって与えられる推定量の漸近的性質を研究した．関連する研究として，Fujisawa and Eguchi (2006)は，β-ダイバージェンスというDPDと同等なダイバージェンスに基づいて，1次元混合正規分布のパラメータを推定するためのアルゴリズムを提案した．また，Fujisawa and Eguchi (2008)では，γ-ダイバージェンスという頑健な推定を考えた．

DPD最小化による推定は，漸近有効ではあるが頑健でないMLEと，頑健ではあるが漸近有効でないL_2推定との間の，架け橋として考えることができる．頑健性と漸近有効性を調節するためには，αの値を適切に決めることが重要である．αの選択法について，Hong and Kim (2001), Warwick and Jones (2005), Warwick (2005), Durio and Isaia (2011)等の研究があるが，モデル選択の立場からは，使い勝手がいいとは必ずしも言い難い．最近，Kobayashi et al. (2011)は，異常値混入モデルにおける潜在分布のクラス数とモデルを有効に推定するためのαの選択法

を提案した．そこでは，異常値混入モデルにおけるDPD (β-ダイバージェンス) を漸近的に評価し，潜在分布のモデル選択のための新しい情報量規準を与えた．

本稿では，Kobayashi et al. (2011) が与えたDPDの漸近理論に基づいて，回帰モデルにおける頑健なモデル選択のための，新しい情報量規準を提案する．提案する情報量規準が，異常値に対して頑健なモデル選択を保証することを，理論的かつ数値的に考察する．

2. 異常値に対して頑健な推定量

d 次元データ $\boldsymbol{x}$ を生成する真の分布の密度関数を $g(\boldsymbol{x})$ とし，その分布関数を $G(\boldsymbol{x})$ とする．$\boldsymbol{\theta} = (\theta_1, ..., \theta_p)^T$ はあるパラメータ数 p をもつパラメータベクトルとする．未知の $g(\boldsymbol{x})$ を近似するモデルを $f(\boldsymbol{x}|\boldsymbol{\theta})$ とし，これを異常値に対して頑健に構築したい．そのために，$f(\boldsymbol{x}|\boldsymbol{\theta})$ の $g(\boldsymbol{x})$ に対する近さを

$$D_\alpha(g(\boldsymbol{x}); f(\boldsymbol{x}|\boldsymbol{\theta})) = \begin{cases} \frac{1}{\alpha}\int_{\boldsymbol{R}^d} g(\boldsymbol{x})^{1+\alpha} d\boldsymbol{x} - \left(1+\frac{1}{\alpha}\right)\int_{\boldsymbol{R}^d} f(\boldsymbol{x}|\boldsymbol{\theta})^\alpha g(\boldsymbol{x}) d\boldsymbol{x} \\ \quad + \int_{\boldsymbol{R}^d} f(\boldsymbol{x}|\boldsymbol{\theta})^{1+\alpha} d\boldsymbol{x} \qquad (\alpha > 0), \\ \int_{\boldsymbol{R}^d} \left(\log g(\boldsymbol{x}) - \log f(\boldsymbol{x}|\boldsymbol{\theta})\right) g(\boldsymbol{x}) d\boldsymbol{x} \quad (\alpha = 0) \end{cases} \tag{2.1}$$

で測る．$D_\alpha(g(\boldsymbol{x}); f(\boldsymbol{x}|\boldsymbol{\theta}))$ はdensity power divergence(DPD) といい，次の性質をもつ．

(i) $D_\alpha(g(\boldsymbol{x}); f(\boldsymbol{x}|\boldsymbol{\theta})) \geq 0$

(ii) $D_\alpha(g(\boldsymbol{x}); f(\boldsymbol{x}|\boldsymbol{\theta})) = 0 \iff g(\boldsymbol{x}) = f(\boldsymbol{x}|\boldsymbol{\theta}),\ \forall \boldsymbol{x} \in \boldsymbol{R}^d$

$D_\alpha(g(\boldsymbol{x}); f(\boldsymbol{x}|\boldsymbol{\theta}))$ が小さいほど，モデル $f(\boldsymbol{x}|\boldsymbol{\theta})$ は真の $g(\boldsymbol{x})$ に近いと考えられる．これ以降は，$\alpha > 0$ の場合を扱う．いま，

$$C_\alpha(g(\boldsymbol{x}); f(\boldsymbol{x}|\boldsymbol{\theta})) = \int_{\boldsymbol{R}^d} f(\boldsymbol{x}|\boldsymbol{\theta})^{1+\alpha} d\boldsymbol{x} - \left(1+\frac{1}{\alpha}\right)\int_{\boldsymbol{R}^d} f(\boldsymbol{x}|\boldsymbol{\theta})^\alpha g(\boldsymbol{x}) d\boldsymbol{x} \tag{2.2}$$

とおく．これを用いて

$$D_\alpha(g(\boldsymbol{x}); f(\boldsymbol{x}|\boldsymbol{\theta})) = C_\alpha(g(\boldsymbol{x}); f(\boldsymbol{x}|\boldsymbol{\theta})) - C_\alpha(g(\boldsymbol{x}); g(\boldsymbol{x})) \tag{2.3}$$

と表せるので，$D_\alpha(g(\boldsymbol{x}); f(\boldsymbol{x}|\boldsymbol{\theta}))$ を最小にするパラメータは，

$$\boldsymbol{\theta}_\alpha = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmin}}\, C_\alpha(g(\boldsymbol{x}); f(\boldsymbol{x}|\boldsymbol{\theta}))$$

で与えられる．ただし，$\boldsymbol{\Theta}$ は $\boldsymbol{\theta}$ の開空間である．(2.2) 式における $g(x)$ の分布関数 $G(x)$ を経験分布関数 $\hat{G}(\boldsymbol{x})$ で置き換え，

$$\ell_\alpha(\boldsymbol{\theta}; f(\boldsymbol{x}|\boldsymbol{\theta})) = \int_{\boldsymbol{R}^d} f(\boldsymbol{x}|\boldsymbol{\theta})^{1+\alpha} d\boldsymbol{x} - \left(1+\frac{1}{\alpha}\right) n^{-1} \sum_{i=1}^{n} f(\boldsymbol{x}_i|\boldsymbol{\theta})^\alpha \tag{2.4}$$

とおく．そのとき，

$$\hat{\boldsymbol{\theta}}_\alpha = \underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmin}}\, \ell_\alpha(\boldsymbol{\theta}; f(\boldsymbol{x}|\boldsymbol{\theta}))$$

を用いて構築したモデル $f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_\alpha)$ は，データ数 n が大きいときに漸近的に最適なモデルと考えられる．$\hat{\boldsymbol{\theta}}_\alpha$ は最小 DPD 推定量 (MDPDE) とよばれ，$\alpha = 0$ のときに最尤推定量 (MLE) と一致し，$\alpha = 1$ のときに Scott (2001) が提案した L_2 推定量と一致する．

いま，$\boldsymbol{I}_\alpha(\boldsymbol{\theta})$, $\boldsymbol{J}_\alpha(\boldsymbol{\theta})$ を

$$\boldsymbol{I}_\alpha(\boldsymbol{\theta}) = \int_{\boldsymbol{R}^d} \frac{\partial \ell_\alpha(\boldsymbol{\theta}; f(\boldsymbol{x}|\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} \frac{\partial \ell_\alpha(\boldsymbol{\theta}; f(\boldsymbol{x}|\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}^T} dG(\boldsymbol{x}),$$

$$\boldsymbol{J}_\alpha(\boldsymbol{\theta}) = \int_{\boldsymbol{R}^d} \frac{\partial^2 \ell_\alpha(\boldsymbol{\theta}; f(\boldsymbol{x}|\boldsymbol{\theta}))}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} dG(\boldsymbol{x})$$

と定義し，$f(\boldsymbol{x}|\boldsymbol{\theta})$ について次の正則条件を仮定する：

(A1) $f(\boldsymbol{x}|\boldsymbol{\theta})$ の台 $A = \{\boldsymbol{x}|f(\boldsymbol{x}|\boldsymbol{\theta}) > 0\}$ は $\boldsymbol{\theta}$ に無関係である．

(A2) $\int_{\boldsymbol{R}^d} f(\boldsymbol{x}|\boldsymbol{\theta})^{1+\alpha} d\boldsymbol{x} < \infty$ であり，すべての $\boldsymbol{x}$ で $f(\boldsymbol{x}|\boldsymbol{\theta})$ は $\boldsymbol{\theta}$ に関して 3 階連続微分可能である．また，$\boldsymbol{\theta}$ に関する微分と $\boldsymbol{x}$ に関する積分の順序交換可能である．

(A3) $\boldsymbol{J}_\alpha(\boldsymbol{\theta}_\alpha) > \boldsymbol{O}$ であり，すべての成分は有界である．

(A4) すべての $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ について

$$\left| \frac{\partial^3 \ell_\alpha(\boldsymbol{\theta}; f(\boldsymbol{x}|\boldsymbol{\theta}))}{\partial \theta_i \partial \theta_j \partial \theta_l} \right| \leq M_{ijl}(\boldsymbol{x}), \quad E_G[M_{ijl}(\boldsymbol{x})] < \infty; \quad i,j,l = 1,...,p$$

となる $M_{ijl}(\boldsymbol{x})$ が存在する．

そのとき，MDPDE は次の漸近正規性をもつ．

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_\alpha - \boldsymbol{\theta}_\alpha) \xrightarrow{d} N_p(\boldsymbol{0}, \boldsymbol{J}_\alpha(\boldsymbol{\theta}_\alpha)^{-1} \boldsymbol{I}_\alpha(\boldsymbol{\theta}_\alpha) \boldsymbol{J}_\alpha(\boldsymbol{\theta}_\alpha)^{-1}) \quad \text{as} \quad n \to \infty. \tag{2.5}$$

次に，MDPDE の異常値に対する頑健性を，シミュレーション実験で検証する．真の分布に，次のような異常値が混入した分布を考えた．

$$g(x) = 0.97\phi(x|0,1) + 0.03\psi(x|4 < x < 8)$$

ここで，$\phi(x|0,1)$ は潜在分布を表し，標準正規分布 $N(0,1)$ を仮定した．一方，$\psi(x|4 < x < 8)$ は異常値の分布を表し，区間 (4,8) の一様分布 $U(4,8)$ を仮定した．つまり，真の分布には，3% の確率で異常値が混入した状況を考えた．真の分布の平均と分散は，$E(x) = 0.180$, $V(x) = 2.058$ である．$g(x)$ から独立にデータを 100 個発生させ，真の分布の平均と分散を MLE ($\alpha = 0$) と MDPDE (α=0.1, 0.3,

0.9) で推定した．これを独立に 1000 回繰り返し，推定値の平均とその分散（括弧内）を纏めたものが表 1 である．MLE は，異常値の影響を忠実に反映して，異常値が混在した真の分布の平均と分散を推定した．それに対して，MDPDE は，異常値の影響を抑えて潜在分布の平均と分散を推定する様子が見て取れた．しかしながら，α の値が小さいとき，MDPDE は異常値の影響を十分に抑えられているわけでなく，また，α の値が大き過ぎると，推定値の分散が大きくなり推定が不安定になることも見て取れた．したがって，MDPDE の使用には，適切な α の値を選択することが重要であるといえる．

表 1. MLE ($\alpha = 0$) と MDPDE (α=0.1, 0.3, 0.9) による
平均と分散の推定値とその分散（括弧内）

$\theta \setminus \alpha$	0	0.1	0.3	0.9
$\mu = 0$	0.197	0.059	-0.014	-0.019
	(0.024)	(0.016)	(0.012)	(0.015)
$\sigma^2 = 1$	2.191	1.418	0.990	0.986
	(0.480)	(0.199)	(0.028)	(0.050)

3. 回帰モデルにおける MDPDE

いま，目的変数を y_i, $d+1$ 次元の説明変数ベクトルを $\boldsymbol{x}_i = (1, x_{i1}, ..., x_{id})^T$ とし，独立な n 個のデータ $\{(y_i, \boldsymbol{x}_i); i = 1, ..., n\}$ を観測したとする．線形回帰モデル

$$y_i = \boldsymbol{\beta}^T \boldsymbol{x}_i + \varepsilon_i \tag{3.1}$$

を考える．ここで，$\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_d)^T$ は $d+1$ 次元の回帰係数ベクトルで，ε_i, $i = 1, ..., n$ は独立同分布に従う誤差変数である．誤差変数 ε の密度関数 $g(\varepsilon)$ は，ある開領域 $\boldsymbol{R}_o$ に確率 $\tau \in (0, 1)$ で異常値が混入した次の分布を仮定する．

$$g(\varepsilon) = (1 - \tau)\phi(\varepsilon|0, \sigma^2) + \tau\psi(\varepsilon) \tag{3.2}$$

ここで，$\phi(\varepsilon|0, \sigma^2)$ は潜在分布となる正規分布 $N(\varepsilon|0, \sigma^2)$ の密度関数を表す．一方，$\psi(\cdot)$ は異常値の密度関数を表し，

$$\int_{\boldsymbol{R}_o} \psi(\varepsilon)d\varepsilon = 1, \quad \sup_{\varepsilon \in \boldsymbol{R}_o} \psi(\varepsilon) < \infty \tag{3.3}$$

を満たすとする．異常値が混入する開領域 $\boldsymbol{R}_o$ における潜在分布の確率

$$\int_{\boldsymbol{R}_o} \phi(\varepsilon|0, \sigma^2)d\varepsilon \ (= \delta > 0, \text{ say})$$

は十分小さいと仮定する．我々は，異常値に頑健な $\boldsymbol{\beta}$ と σ^2 の推定量を MDPDE を用いて構築する．なお，Durio and Isaia (2011) も回帰モデルにおいて MDPDE を考え，α の選択法を議論している．

モデル $f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})$ は，パラメータ $\boldsymbol{\theta}=(\boldsymbol{\beta}^T,\sigma^2)^T$ をもつ正規分布 $N(y_i|\boldsymbol{x}_i;\boldsymbol{\beta}^T\boldsymbol{x}_i,\sigma^2)$

$$f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\Big(-\frac{(y_i-\boldsymbol{\beta}^T\boldsymbol{x}_i)^2}{2\sigma^2}\Big) \tag{3.4}$$

を仮定する．このとき，$\alpha>0$ において

$$\begin{aligned}\ell_\alpha(\boldsymbol{\theta};f(y|\boldsymbol{x};\boldsymbol{\theta}))&=n^{-1}\sum_{i=1}^{n}\int_{\boldsymbol{R}}f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})^{1+\alpha}dy_i-\Big(1+\frac{1}{\alpha}\Big)n^{-1}\sum_{i=1}^{n}f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})^{\alpha}\\&=\frac{1}{\sqrt{(2\pi\sigma^2)^{\alpha}(1+\alpha)}}-\Big(1+\frac{1}{\alpha}\Big)n^{-1}\sum_{i=1}^{n}f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})^{\alpha}\end{aligned} \tag{3.5}$$

と定義して，MDPDE を

$$\hat{\boldsymbol{\theta}}_\alpha=\underset{\boldsymbol{\theta}\in\boldsymbol{\Theta}}{\operatorname{argmin}}\,\ell_\alpha(\boldsymbol{\theta};f(y|\boldsymbol{x};\boldsymbol{\theta}))$$

で求める．なお，$\alpha=0$ においては MLE と一致し，$\boldsymbol{y}=(y_1,...,y_n)^T$，$\boldsymbol{X}=(\boldsymbol{x}_1,...,\boldsymbol{x}_n)$ に対して $\hat{\boldsymbol{\beta}}=(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}\boldsymbol{y}$，$\hat{\sigma}^2=\sum_{i=1}^{n}(y_i-\hat{\boldsymbol{\beta}}^T\boldsymbol{x}_i)^2/n$ とおけば，$\hat{\boldsymbol{\theta}}_0=(\hat{\boldsymbol{\beta}}^T,\hat{\sigma}^2)^T$ となる．

表 2. MLE ($\alpha=0$) と MDPDE (α=0.2, 0.4, 0.6) による
$\boldsymbol{\beta}$ と σ^2 の推定値とその分散（括弧内）

$\alpha\setminus\boldsymbol{\theta}$	$\beta_0=1.0$	$\beta_1=-1.0$	$\beta_2=-0.5$	$\beta_3=0.5$	$\beta_4=1.0$	$\sigma^2=1$
0	1.327	-1.029	-0.467	0.489	0.989	2.545
	(0.623)	(0.152)	(0.150)	(0.145)	(0.131)	(0.946)
0.2	1.063	-1.014	-0.483	0.488	0.983	1.104
	(0.302)	(0.069)	(0.065)	(0.073)	(0.058)	(0.197)
0.4	1.043	-1.015	-0.485	0.485	0.977	0.919
	(0.312)	(0.071)	(0.066)	(0.075)	(0.061)	(0.051)
0.6	1.045	-1.015	-0.485	0.484	0.975	0.888
	(0.354)	(0.080)	(0.076)	(0.084)	(0.070)	(0.057)

表 2 は，異常値に対する MDPDE の頑健性をシミュレーション実験で検証したものである．ここでは，$d=4$ の線形回帰モデル

$$y_i=\beta_0+\sum_{j=1}^{4}\beta_j x_{ij}+\varepsilon_i$$

を考え，$(\beta_0,\beta_1,\beta_2,\beta_3,\beta_4)=(1,-1,-0.5,0.5,1)$ と固定し，$x_{ij},j=1,...,4$ は一様分布 $U(0,2)$ から独立に発生させた．誤差変数 ε は，異常値の混入した以下の分布を考えた．

$$g(\varepsilon)=0.95\phi(\varepsilon|0,1)+0.05\psi(\varepsilon|4<\varepsilon<8)$$

ここで，潜在分布にあたる $\phi(\varepsilon|0,1)$ は標準正規分布 $N(0,1)$ とし，異常値の分布にあたる $\psi(\varepsilon|4<\varepsilon<8)$ は一様分布 $U(4,8)$ とした．$(y_i, \boldsymbol{x}_i)$ なるデータを独立に60個発生させ，回帰係数 $\boldsymbol{\beta}=(1,-1,-0.5,0.5,1)^T$ と潜在分布の分散 $\sigma^2=1$ を，MLE ($\alpha=0$) と MDPDE (α=0.2, 0.4, 0.6) で推定した．この推定を独立に1000回繰り返し，推定値の平均とその分散（括弧内）を計算した．切片 β_0 と分散 σ^2 の推定において，MLEは異常値の影響を顕著に受けた．一方で，MDPDEは異常値の影響を受けることなく，概ね全てのパラメータを良好に推定している様子が見て取れた．

4. 回帰モデルにおける頑健なモデル選択

線形回帰モデル (3.1) において，d 個の説明変数のうちある $k_\star$ ($\leq d$) 個からなる $\boldsymbol{x}_{i(\star)}=(1,x_{i(1\star)},...,x_{i(k_\star\star)})^T$ によって構築される

$$y_i=\boldsymbol{\beta}_{(\star)}^T\boldsymbol{x}_{i(\star)}+\varepsilon_i \tag{4.1}$$

を，真のモデルと仮定する．いま，d 個の説明変数から $\boldsymbol{x}_{i(c)}=(1,x_{i(1)},...,x_{i(k)})^T$ を選択し，真の密度関数を近似する候補モデルとして，パラメータ $\boldsymbol{\theta}_{(c)}=(\boldsymbol{\beta}_{(c)}^T,\sigma_{(c)}^2)^T$ をもつ正規分布 $N(y_i|\boldsymbol{x}_{i(c)};\boldsymbol{\beta}_{(c)}^T\boldsymbol{x}_{i(c)},\sigma_{(c)}^2)$ の密度関数 $f(y_i|\boldsymbol{x}_{i(c)};\boldsymbol{\theta}_{(c)})$ を考える．$\alpha>0$ における $C_\alpha(g(\varepsilon);f(y|\boldsymbol{x}_{(c)};\boldsymbol{\theta}_{(c)}))$ を

$$\begin{aligned}&C_\alpha(g(\varepsilon);f(y|\boldsymbol{x}_{(c)};\boldsymbol{\theta}_{(c)}))\\&=\frac{1}{\sqrt{(2\pi\sigma_{(c)}^2)^\alpha(1+\alpha)}}-(1+\frac{1}{\alpha})n^{-1}\sum_{i=1}^n\int_{\boldsymbol{R}}f(y_i|\boldsymbol{x}_{i(c)};\boldsymbol{\theta}_{(c)})^\alpha g(\varepsilon_i)d\varepsilon_i\end{aligned} \tag{4.2}$$

と定義する．そのとき，

$$\boldsymbol{\theta}_{\alpha(c)}=\underset{\boldsymbol{\theta}_{(c)}\in\boldsymbol{\Theta}}{\operatorname{argmin}}\,C_\alpha(g(\varepsilon);f(y|\boldsymbol{x}_{(c)};\boldsymbol{\theta}_{(c)}))$$

とおき，これのMDPDEを (3.5) 式から

$$\hat{\boldsymbol{\theta}}_{\alpha(c)}=\underset{\boldsymbol{\theta}_{(c)}\in\boldsymbol{\Theta}}{\operatorname{argmin}}\,\ell_\alpha(\boldsymbol{\theta}_{(c)};f(y|\boldsymbol{x}_{(c)};\boldsymbol{\theta}_{(c)}))$$

によって与える．MDPDEについて，Kobayashi et al. (2011) の定理4.1と同様に，次の定理が成り立つ．

定理1. もしも $\boldsymbol{x}_{i(c)}$ が真のモデルの説明変数 $x_{i(1\star)},...,x_{i(k_\star\star)}$ をすべて含むならば，$\tau\to0$, $\delta\to0$ のもとで

$$D_\alpha(g(\varepsilon);f(y|\boldsymbol{x}_{(c)};\boldsymbol{\theta}_{\alpha(c)}))=O(\tau^2)+O(\tau^{1+\alpha})+O(\delta^{1+\alpha})$$

が成り立つ．

注意1. 真のモデルの説明変数 $x_{i(1\star)},...,x_{i(k_\star\star)}$ の中で $\boldsymbol{x}_{i(c)}$ に含まれないものが存在するとき，次を仮定をする．

$$\boldsymbol{\beta}_{\alpha(c)}^T\boldsymbol{x}_{i(c)} \neq \boldsymbol{\beta}_{(\star)}^T\boldsymbol{x}_{i(\star)},\ i=1,...,n$$

そのとき，$D_\alpha(g(\varepsilon);f(\boldsymbol{y}|\boldsymbol{x}_{(c)};\boldsymbol{\theta}_{\alpha(c)}))>0$ となる．

いま，

$$\ell_\alpha(\boldsymbol{\theta};f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})) = \frac{1}{\sqrt{(2\pi\sigma^2)^\alpha(1+\alpha)}} - \left(1+\frac{1}{\alpha}\right)f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})^\alpha$$

とおく．それに伴い，

$$\boldsymbol{I}_\alpha(\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^n\int_{\boldsymbol{R}}\frac{\partial\ell_\alpha(\boldsymbol{\theta};f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta}))}{\partial\boldsymbol{\theta}}\frac{\partial\ell_\alpha(\boldsymbol{\theta};f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta}))}{\partial\boldsymbol{\theta}^T}dG(\varepsilon_i),$$

$$\boldsymbol{J}_\alpha(\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^n\int_{\boldsymbol{R}}\frac{\partial^2\ell_\alpha(\boldsymbol{\theta};f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta}))}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}dG(\varepsilon_i)$$

を定義する．そのとき，Kobayashi et al. (2011) の定理4.2と同様に，次の定理が成り立つ．

定理2. $\tau=o(n^{-1/2})$ と仮定する．もしも $\boldsymbol{x}_{i(c)}$ が真のモデルの説明変数 $x_{i(1\star)},...,x_{i(k_\star\star)}$ をすべて含むならば，$\tau^{1+\alpha}=o(n^{-1})$，$\delta^{1+\alpha}=o(n^{-1})$ のもとで

$$D_\alpha(g(\varepsilon);f(y|\boldsymbol{x}_{(c)};\hat{\boldsymbol{\theta}}_{\alpha(c)})) = (2n)^{-1}\mathrm{tr}\{\boldsymbol{I}_\alpha(\boldsymbol{\theta}_{\alpha(c)})\boldsymbol{J}_\alpha(\boldsymbol{\theta}_{\alpha(c)})^{-1}\}+o(n^{-1})$$

が成り立つ．

2節で見たように，MDPDEは α の値を上手く選択することによって，MLEにおける異常値の影響を改善した．モデル選択においても，DPDに基づく情報量規準が，Akaike (1973) のAICやTakeuchi (1976) のTICなどのK-L情報量に基づく情報量規準を，異常値の影響について改善することが期待される．いま，

$$\hat{\boldsymbol{I}}_\alpha(\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^n\frac{\partial\ell_\alpha(\boldsymbol{\theta};f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta}))}{\partial\boldsymbol{\theta}}\frac{\partial\ell_\alpha(\boldsymbol{\theta};f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta}))}{\partial\boldsymbol{\theta}^T},$$

$$\hat{\boldsymbol{J}}_\alpha(\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^n\frac{\partial^2\ell_\alpha(\boldsymbol{\theta};f(y_i|\boldsymbol{x}_i;\boldsymbol{\theta}))}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}$$

とおき，$E_G(\mathrm{tr}\{\hat{\boldsymbol{I}}_\alpha(\hat{\boldsymbol{\theta}}_{\alpha(c)})\hat{\boldsymbol{J}}_\alpha(\hat{\boldsymbol{\theta}}_{\alpha(c)})^{-1}\}) = \mathrm{tr}\{\boldsymbol{I}_\alpha(\boldsymbol{\theta}_{\alpha(c)})\boldsymbol{J}_\alpha(\boldsymbol{\theta}_{\alpha(c)})^{-1}\}+o(1),\ n\to\infty$ を仮定する．

[**DPD 最小化モデル選択基準**] 次で定義される $\mathrm{IC}_{\alpha(c)}$ が最小のモデルを，予測の意味で潜在分布に近いモデルとして選択する．

$$\mathrm{IC}_{\alpha(c)} = 2n\ell_\alpha(\hat{\boldsymbol{\theta}}_{\alpha(c)}; f(y|\boldsymbol{x}_{(c)}; \hat{\boldsymbol{\theta}}_{\alpha(c)})) + 2\mathrm{tr}\{\hat{\boldsymbol{I}}_\alpha(\hat{\boldsymbol{\theta}}_{\alpha(c)})\hat{\boldsymbol{J}}_\alpha(\hat{\boldsymbol{\theta}}_{\alpha(c)})^{-1}\} \tag{4.3}$$

もしも，$k+1$ 個の説明変数をもつ $\boldsymbol{x}_{i(c)}$ が真のモデルの説明変数をすべて含み，$k+2$ 個の説明変数をもつ $\boldsymbol{x}_{i(c')}$ が $\boldsymbol{x}_{i(c)}$ の説明変数をすべて含むとき，次の条件を仮定する．

$$0 < \mathrm{tr}\{\boldsymbol{I}_\alpha(\boldsymbol{\theta}_{\alpha(c)})\boldsymbol{J}_\alpha(\boldsymbol{\theta}_{\alpha(c)})^{-1}\} < \mathrm{tr}\{\boldsymbol{I}_\alpha(\boldsymbol{\theta}_{\alpha(c')})\boldsymbol{J}_\alpha(\boldsymbol{\theta}_{\alpha(c')})^{-1}\}$$

そのとき，Kobayashi et al. (2011) の定理 4.3 と同様に，次の定理を得る．

定理 3. $\tau^2 = o(n^{-1})$, $\tau^{1+\alpha} = o(n^{-1})$, $\delta^{1+\alpha} = o(n^{-1})$ を仮定する．そのとき，真のモデルの $\mathrm{IC}_{\alpha(\star)}$ の期待値 $E_G[\mathrm{IC}_{\alpha(\star)}]$ は，候補モデル $N(y_i|\boldsymbol{x}_{i(c)}; \boldsymbol{\beta}_{(c)}^T\boldsymbol{x}_{i(c)}, \sigma_{(c)}^2)$ の $\mathrm{IC}_{\alpha(c)}$ の期待値 $E_G[\mathrm{IC}_{\alpha(c)}]$ を最小にする．

注意 2. α の値は，定理 3 の収束条件を満たす範囲で選択する．真のモデルの情報量基準 $\mathrm{IC}_{\alpha(\star)}$ が平均的に最小値を与えることから，DPD 最小化モデル選択基準によって真のモデルを平均的に選択することができる．

5. シミュレーション

4 節で提案した情報量基準 $\mathrm{IC}_{\alpha(c)}$ の性能を，シミュレーション実験で検証する．次のような線形回帰モデルを考えた．

$$y_i = \beta_0 + \sum_{j=1}^{6} \beta_j x_{ij} + \varepsilon_i$$

ここで，$x_{ij}, j = 1, ..., 6$ は，一様分布 $U(0, 2)$ から独立に発生させた．誤差変数 ε_i は，異常値が混入した以下の分布を考えた．

$$g(\varepsilon) = 0.95\phi(\varepsilon|0, 1) + 0.05\psi(\varepsilon|4 < \varepsilon < 8)$$

真のモデルに，次の 2 つのモデルを考えた．

(I) $\beta_0 = 1,\ \beta_1 = -1,\ \beta_2 = 1,\ \beta_3 = \cdots = \beta_6 = 0$

(II) $\beta_0 = 1,\ \beta_1 = -1,\ \beta_2 = -0.5,\ \beta_3 = 0.5,\ \beta_4 = 1,\ \beta_5 = \beta_6 = 0$

候補モデル $N(y_i|\boldsymbol{x}_{i(c)}; \boldsymbol{\beta}_{(c)}^T\boldsymbol{x}_{i(c)}, \sigma_{(c)}^2)$ については，$\boldsymbol{x}_{i(c)}(= \boldsymbol{x}_{ik(c)}) = (1, x_{i1}, ..., x_{ik})^T$, $k = 1, ..., 6$ の 6 個のモデルを考えた．真のモデルは，(I) の場合は $\boldsymbol{x}_{i2(c)}$，(II) の場合は $\boldsymbol{x}_{i4(c)}$ で与えられることに注意する．いま，真の分布から独立にデータを 60 $(= n)$ 個発生させ，潜在分布の推定を，MLE $(\alpha = 0)$ と MDPDE $(\alpha = 0.2, 0.4, 0.6)$ で

行った．構築した6個のモデルの中から，TIC ($\alpha = 0$) と $\mathrm{IC}_{\alpha(c)}$ ($\alpha = 0.2, 0.4, 0.6$) を用いて，各 α について情報量規準を最小にするモデルを選択した．この実験を独立に1000回繰り返し，TIC ($\alpha = 0$) と $\mathrm{IC}_{\alpha(c)}$ ($\alpha = 0.2, 0.4, 0.6$) による各モデルの選択回数を，表3に纏めた．

表3. TIC ($\alpha = 0$) と $\mathrm{IC}_{\beta(c)}$ ($\alpha = 0.2, 0.4, 0.6$) による各モデルの選択回数（実験回数：1000回）

真のモデルが (I) の場合

$\alpha \setminus k$	1	2	3	4	5	6
0	160	494	155	74	63	54
0.2	16	631	179	85	66	23
0.4	11	628	192	91	53	25
0.6	24	645	203	79	37	12

真のモデルが (II) の場合

$\alpha \setminus k$	1	2	3	4	5	6
0	142	71	35	483	149	120
0.2	33	15	4	713	153	82
0.4	18	12	3	733	157	77
0.6	37	28	17	690	172	56

MLEに基づくTICは異常値に敏感で，真のモデルを有効に選択していないことが見て取れた．一方，MDPDEに基づく $\mathrm{IC}_{\beta(c)}$ は，真のモデルをTICよりも高い確率で選択していることが見て取れた．設定を変えた実験においても同様の結果が得られ，DPD最小化モデル選択基準 $\mathrm{IC}_{\beta(c)}$ が，異常値に対して頑健なモデル選択を実現していることが確認できた．

謝辞　本研究は，科学研究費補助金 基盤研究 (B) 22300094 研究代表者: 青嶋 誠「高次元データの理論と方法論の総合的研究」から，研究助成を受けています．

参考文献

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *2nd Inter. Symp. on Information Theory* (Petrov, B. N. and Csaki, F., eds.), Akademiai Kiado, 267–281 (Reproduced in *Breakthroughs in Statistics*, 1 (Kotz, S. and Johnson, N. L., eds.), Springer-Verlag (1992)).

Basu, A., Harris, I. R., Hjort, N. L. and Jones, M. C. (1998). Robust and efficient estimation by minimising a density power divergence. *Biometrika* **85**, 549-559.

Basu, S., Basu, A. and Jones, M. C. (2006). Robust and efficient parametric estimation for censored survival data. *Ann. Inst. Statist. Math.* **58**, 341-355.

Durio, A. and Isaia, E.D. (2011). The minimum density power divergence approach in building robust regression models. *Informatica* **22**, 43-56.

Fujisawa, H. and Eguchi, S. (2006). Robust estimation in the normal mixture model. *J. Statist. Plan. Infer.* **136**, 3989–4011.

Fujisawa, H. and Eguchi, S. (2008). Robust parameter estimation with a small bias against heavy contamination. *J. Multivariate Anal.* **99**, 2053–2081.

Hjort, N. L. (1994). Minimum L2 and robust Kullback-Leibler estimation. *Proceedings of the 12th Prague Conference on Information Theory, Statistical Decision Functions and Random Process* (P. Lachout and J. Visek, eds.), 102–105. Prague: Academy of Sciences of the Czech Republic.

Hong, C. and Kim, Y. (2001). Automatic selection of the tuning parameter in the minimum density power divergence estimation. *Journal of the Korean Statistical Association.* **30**, 453–465.

Jones, M. C., Hjort, N. L., Harris, I. R. and Basu, A. (2001). A comparison of related density-based minimum divergence estimators. *Biometrika* **88**, 865–873.

Kobayashi, Y., Yata, K. and Aoshima, M. (2011). Robust and accelerated model selection for clustering in a contaminated mixture model, submitted.

Scott, D. W. (2001). Parametric statistical modeling by minimum integrated square error. *Technometrics* **43**, 274–285.

Takeuchi, K. (1976). Distribution of information statistics and criteria for adequacy of models. *Mathematical Science* **153** 12-18 (in Japanese).

Warwick, J. (2005). A data-based method for selecting tuning parameters in minimum distance estimators. *Computational Statistics and Data Analysis* **48**, 571–585.

Warwick, J. and Jones, M. C. (2005). Choosing a robustness tuning parameter. *Journal of Statistical Computation and Simulation* **75**, 581–588.